applied sciences

Article

Three-Dimensional Human Pose Estimation with Spatial–Temporal Interaction Enhancement Transformer

Haijian Wang [1,2], Qingxuan Shi [1,2,*] and Beiguang Shan [1,2]

[1] School of Cyber Security and Computer, Hebei University, Baoding 071002, China
[2] Hebei Machine Vision Engineering Research Center, Hebei University, Baoding 071002, China
* Correspondence: qingxuanshi@hbu.edu.cn

Abstract: Three-dimensional human pose estimation is a hot research topic in the field of computer vision. In recent years, significant progress has been made in estimating 3D human pose from monocular video, but there is still much room for improvement in this task owing to the issues of self-occlusion and depth ambiguity. Some previous work has addressed the above problems by investigating spatio-temporal relationships and has made great progress. Based on this, we further explored the spatio-temporal relationship and propose a new method, called STFormer. Our whole framework consists of two main stages: (1) extract features independently from the temporal and spatial domains; (2) modeling the communication of information across domains. The temporal dependencies were injected into the spatial domain to dynamically modify the spatial structure relationships between joints. Then, the results were used to refine the temporal features. After the preceding steps, both spatial and temporal features were strengthened, and the estimated final pose will be more precise. We conducted substantial experiments on a well-known dataset (Human3.6), and the results indicated that STFormer outperformed recent methods with an input of nine frames. Compared to PoseFormer, the performance of our method reduced the MPJPE by 2.1%. Furthermore, we performed numerous ablation studies to analyze and prove the validity of the various constituent modules of STFormer.

Keywords: three-dimensional human pose estimation; temporal dependencies; spatio-temporal relationships; cross-domain information interaction

check for updates

Citation: Wang, H.; Shi, Q.; Shan, B. Three-Dimensional Human Pose Estimation with Spatial–Temporal Interaction Enhancement Transformer. *Appl. Sci.* **2023**, *13*, 5093. https://doi.org/10.3390/app13085093

Academic Editors: Flavio Bertini, Eleonora Iotti and Vincenzo Bonnici

Received: 7 March 2023
Revised: 16 April 2023
Accepted: 17 April 2023
Published: 19 April 2023

1. Introduction

Three-dimensional human pose estimation is a task that predicts the 3D spatial position of human joints from images or videos, which has broad application fields such as action recognition [1–4], human–computer interaction [5], augmented reality [6], and autonomous driving [7]. Many advanced approaches [8–11] solve this task by decoupling it into two subtasks, i.e., first locating 2D keypoint coordinates with a 2D pose detector and then designing a 2D-to-3D lifting network to infer the joint positions in 3D space from the 2D keypoints. Despite their impressive performance, it is still a tricky problem as different poses in 3D space may have the same pose when projected to 2D due to depth ambiguity. Many works [12,13] have made significant progress in addressing this issue by employing multiple simultaneous cameras to view objects from various viewpoints. However, in contrast to monocular methods, multi-view methods have strict prerequisites for equipment and the environment, which are not practical in reality. Therefore, many approaches are beginning to explore spatial and temporal information in monocular videos [8,9,14,15]. These models take video as the input and perceive the depth information of moving objects using the temporal information of the sequence.

Currently, convolutional neural networks (CNNs) are widely applied to various tasks in computer vision [16–21], and all of them have achieved relatively favorable performance. Unlike CNNs, graph convolutional neural networks (GCNs) are more suitable for processing graph structured data. The GCN describes the spatial relationships between joints by a hand-crafted graph adjacency matrix. The graph is built on an articulated human skeleton, where the human joints represent nodes and the human skeleton forms edges. Cai et al. [15] designed a spatio-temporal graph model based on a graph convolutional neural network to estimate 3D joint positions from 2D pose sequences using spatio-temporal relations. However, the GCN-based approach has two limitations: First, each node in the graph shares a transformation matrix. This weight sharing prevents the GCN from learning different patterns of relationships between different body joints. However, for different poses, the relationship between their joints is different [22]. For example, in the running pose, there is a close relationship between the hands and the feet, while for the sitting pose, this is not the case. Such information is difficult to capture with a static skeleton graph. Liu et al. [23] solved this problem by weight separation and performing different feature transformations on different nodes before aggregating the features. However, this significantly increases the size of the model. Second, GCN-based methods with small temporal receptive fields prevent modeling long-term dependencies between temporal sequences. In recent years, the Vision Transformer has seen widespread adoption across a broad range of computer vision tasks. Its internal self-attention mechanism allows for flexible modeling of long-range globally consistent information about the input sequence. Zheng et al. [8] proposed a novel approach for 3D human pose estimation from video using a Transformer-based spatio-temporal network that does not rely on convolutional architectures. It first models the intrinsic structural relationships between joints in the spatial domain and then acquires the temporal consistency of the video sequences. However, this sequential network only models the static spatial relationship of each frame, but ignores the effect of temporal information on the spatial structure.

Based on this, we designed a Transformer-based spatial–temporal interaction enhancement network (STFormer) for estimating 3D body poses from monocular videos. It adopts the interaction of different domain features to enhance the representation of the current domain, which is crucial for accurately predicting the position of body joints. To accomplish this, STFormer starts with the generation of coarse temporal and spatial representations and then continuously communicates between them to eventually produce a more accurate 3D prediction. This framework more effectively extracts spatial and temporal features and also builds stronger connections between them. Specifically, in the first stage, we propose the the Feature Extraction (FE) module, which consists of two branches: Spatial Feature Extraction (SFE) branch and Temporal Feature Extraction (TFE) branch, which are responsible for extracting the intrinsic spatial structure of each frame and the temporal dependencies between frames. Although spatial and temporal features are extracted in the first stage, there is no information interaction between them. In view of the influence of temporal information on the spatial structure as discussed in the above paragraph, in the second stage, the connection between spatial and temporal information is established through the Cross-Domain Interaction (CDI) module. It consists of two blocks: the Spatial Reconstruction (SR) block, which is responsible for injecting temporal features into the spatial domain, and the Temporal Refinement (TR) block, which uses the reconstructed features to refine temporal features. CDI captures mutual spatio-temporal correlations to construct cross-domain communication, enabling messages to be passed between spatial and temporal representations for better interaction modeling.

With the proposed STFormer, temporal and spatial features are explicitly incorporated into the Transformer model. The spatial structure of each video frame is dynamically changed according to the video sequence information, and then, this adjusted spatial structure, in turn, contributes to the temporal representation. As a result, both representations are significantly enhanced to provide poses that are more accurate. The summarization of our contributions are as follows:

1. To predict 3D human pose more accurately from monocular videos, we designed a spatio-temporal interaction enhanced Transformer network, called STFormer. STFormer is a two-stage method, in which the first stage extracts features independently from

the spatial and temporal domains, respectively, and the second stage interacts spatial and temporal information across domains to enrich the representations.

2. In the second stage, we designed the Spatial Reconstruction block and the Temporal Refinement block. The Spatial Reconstruction block injects the temporal features into the spatial domain to adjust the spatial structure relationship; the reconstructed features are then sent to the Temporal Refinement block to complement the weaker intra-frame structural information in the temporal features.

2. Related Work

2.1. Three-Dimensional Human Pose Estimation

There are two methods used for 3D human pose estimation, one of which is to predict 3D coordinates directly from the input image, known as the end-to-end method [24–27]. Another is a 2D to 3D lifting technique, which utilizes an available 2D pose detector to generate 2D coordinates and then lifts them to the 3D pose through a lifting network [8,9,15,28,29]. However, the end-to-end approach takes RGB images as the input, which requires high computational consumption and is impractical in real-world applications [11]. In contrast, the inputs for lifting methods are 2D coordinates, which are simpler to process than RGB images. For example, Reference [29] suggested a residual network that is fully connected and lifts the keypoints of each frame from the 2D to the 3D space. Recent work [30] has shown that the lifting method can efficiently and accurately regress 3D poses utilizing detected 2D keypoints, thanks to the superior performance of 2D human pose estimation, and this method outperforms end-to-end approaches. As in [29,30], our proposed model belongs to the category of 2D to 3D lifting.

2.2. Three-Dimensional Human Pose Estimation for Video under Lifting Method

Some previous work has been devoted to single-frame 3D human pose estimation. Recently, in order to improve the accuracy of pose estimation and reduce depth ambiguity, many approaches have tried to exploit the temporal information of video sequences [9,10,14,15,28]. Hossain and Little [28] designed an LSTM-based model to predict three-dimensional poses using temporally consistent information from the input sequence. Pavllo et al. [14] presented VideoPose3D, which applies dilated temporal convolution on 2D joint sequences to capture long-term information for predicting 3D poses. Chen et al. [10] proposed a new solution that decouples the video 3D human pose estimation task into predicting bone length and orientation, rather than directly regressing the position of 3D joints. Our proposed method considers both the structural relationships between joints in monocular images and the temporal consistency of video sequences.

2.3. Visual Transformer

Vaswani et al. [31] first proposed Transformer with a powerful self-attention mechanism and applied it to natural language processing tasks. Since then, Transformer has been gradually extended to the field of computer vision and has shown good performance in various vision tasks [32–34]. Dosovitskiy et al. [35] proposed a fundamental Transformer structure with state-of-the-art picture classification capabilities. Carion et al. [36] introduced DETR, a new paradigm for Transformer-based end-to-end object detection. Regarding the human pose estimation, Li et al. [37] proposed a Transformer-based 2D human pose model that represents each keypoint of the human body in terms of tokens. PoseFormer [8] is the first pure Transformer model for monocular video 3D human pose estimation. To reduce the sequence redundancy caused by the high similarity of poses between adjacent frames, Li et al. [11] proposed StrideFormer by replacing the fully connected layer in FFN with 1D stride temporal convolution. Our approach was based on Transformer, which maps 2D keypoints to 3D poses. Unlike these previously mentioned methods, our proposed approach not only extracts representations in the temporal and spatial domains, but also enables cross-domain information exchange to enhance spatial and temporal representations for generating more accurate 3D poses.

3. Method

Figure 1 illustrates the whole framework of our STFormer. We adopted the 2D to 3D lifting approach used in [11,14,29], taking 2D video pose sequences as the input and predicting the 3D pose for intermediate frames. The presented two-stage network STFormer consists of Feature Extraction (FE) and Cross-Domain Interaction (CDI). Specifically, the FE contains two branches, Spatial Feature Extraction (SFE) branch and Temporal Feature Extraction (TFE) branch, which are responsible for modeling the inherent structural of human joints and temporal dependence, respectively. CDI includes a Spatial Reconstruction (SR) block and a Temporal Refinement (TR) block, which are responsible for the interaction of the information extracted in the previous stage.

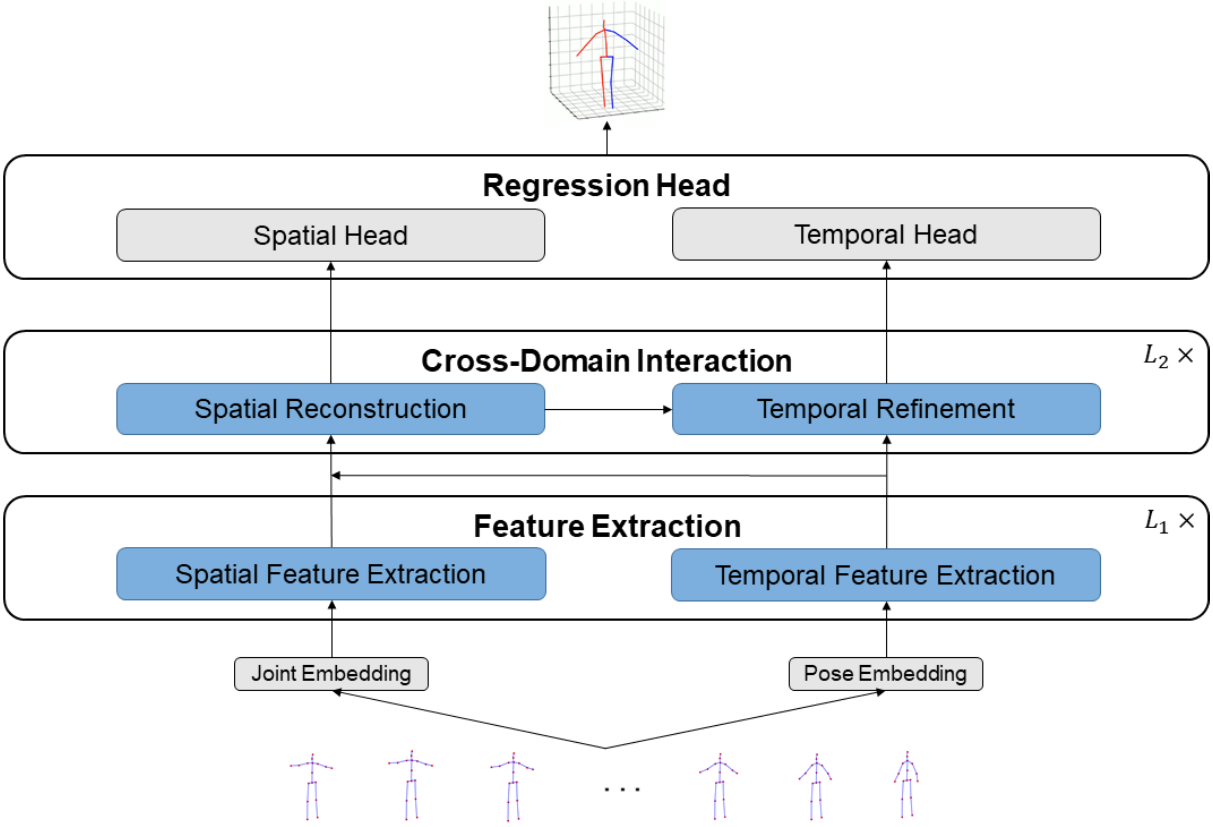

Figure 1. The pipeline of our proposed framework (STFormer) for video 3D human pose estimation. Feature Extraction module for independently extracting features from spatial and temporal domains. The Cross-Domain Interaction module then receives the spatial and temporal features extracted in the previous stage for feature enhancement. Two regression heads to regress the 3D poses in the spatial and temporal domains, respectively.

3.1. Preliminary

Due to the excellent performance of Transformer across tasks, our model also adopted the Transformer-based architecture. In this part, we briefly describe the components of Transformer [31], including scaled dot-product attention, multi-head self attention, and multi-layer perceptron.

Scaled dot-product attention is shown in Figure 2a, and its input consists of queries, keys, and values. First, the dot-product of the query and all the keys are calculated and scaled, then the Softmax function is applied to obtain the weights and finally multiplied with the values to obtain the output. The process is expressed formulaically as follows:

$$SA(Q,\ K,\ V) \ = \ \text{Softmax}(\frac{Q \cdot K^T}{\sqrt{d}}) \cdot V. \tag{1}$$

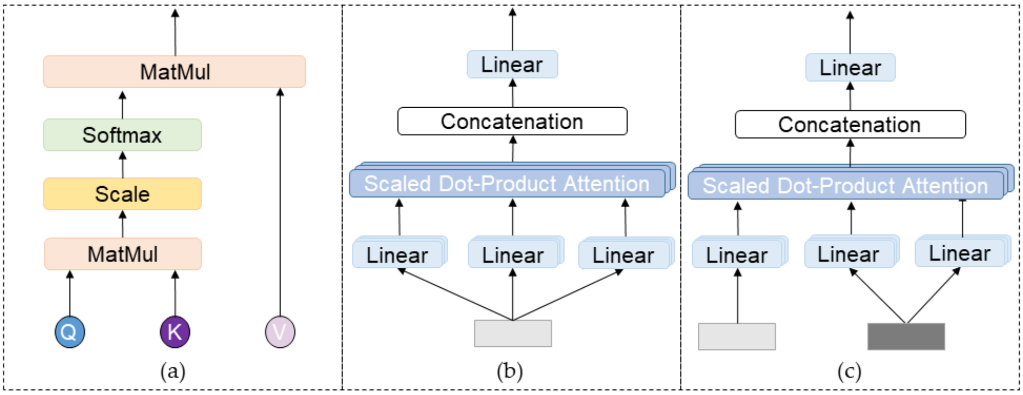

Figure 2. Structure of attention mechanism. (**a**) Scaled Dot-Product Attention. (**b**) Multi-head self-attention (MSA). (**c**) Multi-head cross-attention (MCA).

To enhance the expressive capacity of the network, a multi-head self attention (MSA) mechanism is further proposed. MSA divides queries, keys, and values into h heads, with each head performing scaled dot-product attention in parallel (Figure 2b). The information in different representation subspaces at different locations is jointly modeled with multiple heads. After that, the values that are generated by the output of each head are aggregated and linearly projected to generate the final output.

$$MSA = f(Concat(head_1, head_2, \ldots, head_h)), head_i = SA(Q_i,\ K_i,\ V_i), \tag{2}$$

where $i \in [1, ..., h]$ is the index of the head and $f(\cdot)$ is the linear function.

Multi-layer perceptron (MLP) consists of two linear layers and an activation function for non-linearity and feature transformation. The process is defined as

$$MLP(x) = \sigma(xW_1 + b_1)W_2 + b_2, \tag{3}$$

where W_1 and W_2 are the weights of the two linear layers, b_1 and b_2 are the bias terms, and σ denotes the GELU activation function [38].

3.2. Feature Extraction

In the first stage of STFormer, the input is a 2D pose sequence $X \in \mathbb{R}^{N\times J\times 2}$ with N video frames and J joints per pose, where 2 represents the joint position coordinates, which is pre-processed and fed to the Spatial Feature Extraction branch and the Temporal Feature Extraction branch to extract features, respectively.

3.2.1. Spatial Feature Extraction Branch

For the spatial domain, we propose SFE composed of the Transformer encoder to model the intrinsic structural relationships of the human joints for each frame (see Figure 3a,c). More specifically, we treated each joint of a 2D pose as a token and took a linear projection layer to embed it into high-dimension space: $X^i \in \mathbb{R}^{J\times 2} \xrightarrow{embed} \overline{X^{Si}} \in \mathbb{R}^{J\times C}$, where X^i is the i-th frame and C is the joint embedding dimension. Then, to save the spatial location information, we added a learnable spatial position embedding $E^S_{pos} \in \mathbb{R}^{J\times C}$ with this high-dimensional feature, and it became $Z^{Si}_0 \in \mathbb{R}^{J\times C}$. Finally, $Z^{Si}_0 \in \mathbb{R}^{J\times C}$ were fed into the SFE to extract spatial structural information across all joints. These processes can be described as:

$$\begin{gathered} Z^{Si}_0 = \overline{X^{Si}} + E^S_{pos}, \\ Z^{Si'}_l = Z^{Si}_{l-1} + MSA(LN(Z^{Si}_{l-1})), \\ Z^{Si}_l = Z^{Si'}_l + MLP(LN(Z^{Si'}_l)), \end{gathered} \tag{4}$$

where $LN(\cdot)$ denotes the layer normalization layer and $l \in [1, \cdots, L_1]$ is the index of the FE layers. After going through the L_1-layer FE module, the output of frame i on the SFE branch will be $Z^{Si}_{L_1}$.

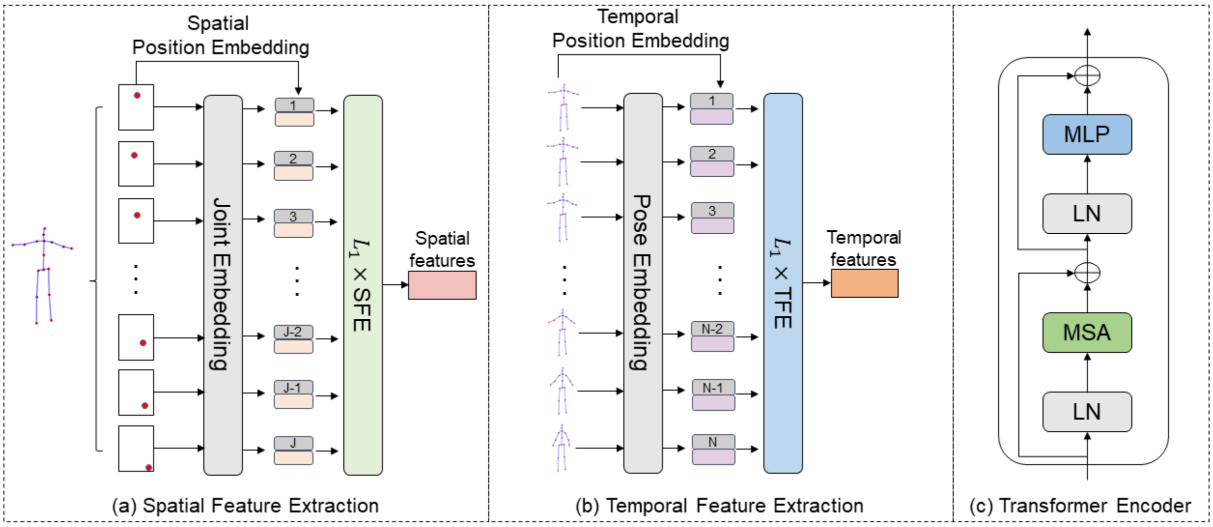

Figure 3. Feature Extraction module. (**a**) Within each frame, the Spatial Feature Extraction (SFE) branch is responsible for extracting information on the intrinsic structure of the body joints. (**b**) Temporal Feature Extraction (TFE) branch for learning the consistency relationships of the whole sequence. (**c**) The description of the Transformer encoder structure.

3.2.2. Temporal Feature Extraction Branch

Although SFE can obtain the intrinsic structure between joints, whereas it ignores the temporal dependencies between video frames. To explore temporally consistent information, we propose an SFE-like branch TFE (see Figure 3b,c). We should construct a feature representation in the temporal domain. To do this, different from the operation in the spatial domain, we treated each frame of the 2D pose as a token in the temporal domain. Then, sent it to the TFE branch to learn the global relationships between the input sequences. For the input $X \in \mathbb{R}^{N \times J \times 2}$, we first combined the coordinates of all joints in each frame, denoted as $X^T \in \mathbb{R}^{N \times (J \cdot 2)}$. Then, as with the initial spatial features, we embedded them into a high-dimensional feature $\overline{X^T} \in \mathbb{R}^{N \times D}$ (D indicates the embedding dimension per frame) and added a learnable position encoding $E^T_{pos} \in \mathbb{R}^{N \times D}$ to retain the frame position information. Finally, we fed the embedded features into the TFE. These procedures can be expressed as:

$$\begin{aligned} \overline{X^T} &= embed(X^T), \\ Z^T_0 &= \overline{X^T} + E^T_{pos}, \\ Z^{T'}_l &= Z^T_{l-1} + MSA(LN(Z^T_{l-1})), \\ Z^T_l &= Z^{T'}_l + MLP(LN(Z^{T'}_l)), \end{aligned} \tag{5}$$

where $l \in [1, \cdots, L_1]$ is the index of the FE layers. After the L_1-layer FE, the output of the TFE branch is $Z^T_{L_1} \in \mathbb{R}^{N \times D}$.

3.3. Cross-Domain Interaction

The structural relationships between the joints of the body are not invariable for different states of motion [22]. For example, for the “running” pose, there is a strong connection between the hands and feet: reaching the left hand is accompanied by a corresponding step of the right foot, whereas for other poses, there may not be such a relationship. Therefore, the corresponding spatial structure relationships should be modeled for different poses. In the temporal domain, we treated each frame pose as a token to extract temporal information, however, ignoring the spatial structure relationship inside each frame. In order to tackle the above issues, a Cross-Domain Interaction module consisting of the Spatial Reconstruction block (SR) and Temporal Refinement block (TR) is proposed. The SR block injects the temporal action information into the spatial domain, allowing the network to adjust the relationship between joints based on this information. Then, the output of the SR block is converted to the temporal domain as the input to the TR block to complement the lack of spatial structure information in the temporal features. For the convenience of formula writing, we use $\widehat{Z_l^{Si}}$, $\widehat{Z_l^T}$ to denote spatial and temporal features, respectively, in the Cross-Domain Interaction module, $\widehat{Z_0^{Si}} = Z_{L_1}^{Si}$ and $\widehat{Z_0^T} = Z_{L_1}^T$.

3.3.1. Spatial Reconstruction

In the first stage, we extracted features from the temporal and spatial domains, respectively. However, both types of information are extracted independently, and there is no interaction between them. To employ temporal action information to reconstruct the joint structural relationship, we need to inject the temporal features into the spatial domain. Before that, we must realize that the temporal and spatial features are inconsistent. In the temporal domain, each frame of a 2D pose sequence is represented by a token whose dimension is $1 \times D$, while in the spatial domain, the feature dimension is $J \times C$, and each token represents one joint. This results in the information from different domains not being able to be directly interacted. To achieve cross-domain temporal information injection, we propose the SR block (see Figure 4a), which consists of the feature transforming unit (FTU), multi-head cross-attention (MCA), MSA, and MLP.

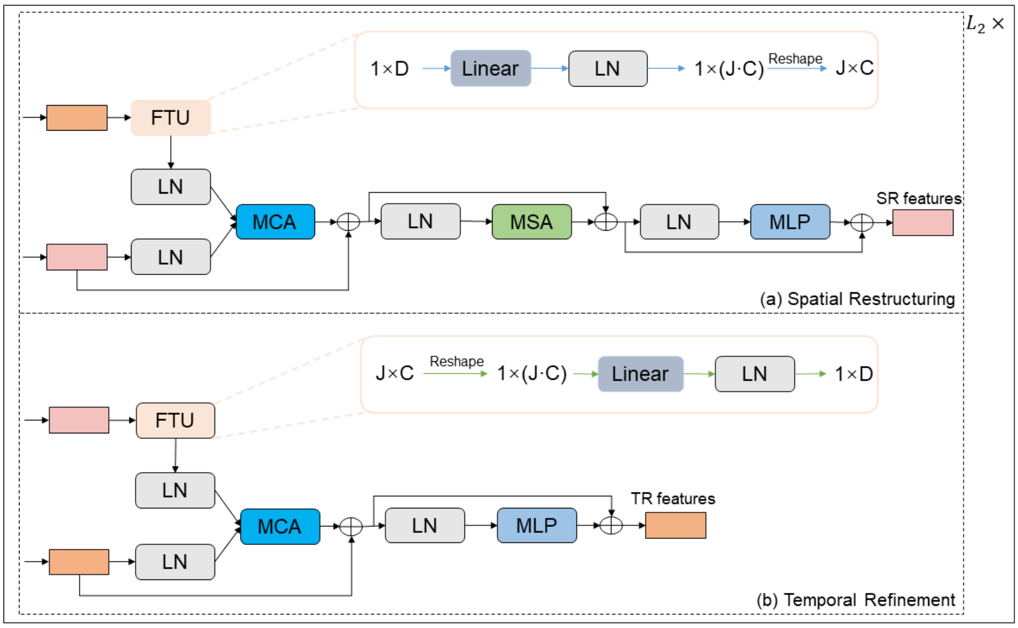

Figure 4. Cross-Domain Interaction. (**a**) The Spatial Reconstruction (SR) block injects the temporal features into the spatial domain after transformation to adjust the spatial structure. (**b**) The Temporal Refinement (TR) block receives the reconstructed spatial features to further refine the temporal features.

We first utilized the FTU to convert the temporal features to the spatial domain. Specifically, a linear layer converts the number of temporal feature channels to $J \cdot C$ and then passes through a LayerNorm layer. Finally, the channels are grouped and reshaped into $J \times C$. After the above operations, the temporal features are successfully transformed from the temporal domain to the spatial domain ($\widehat{Z_l^T} \rightarrow \{\widehat{Z_l^{T\rightarrow S1}}, \widehat{Z_l^{T\rightarrow S2}}, \cdots, \widehat{Z_l^{T\rightarrow SN}}\}$). Next, $\widehat{Z_l^{T\rightarrow Si}}$ and spatial feature $\widehat{Z_l^{Si}}$ are fed to the MCA. The MCA interacts the information from different domains and has a similar structure to the MSA (see Figure 2c). The feature $\widehat{Z_l^{Si}}$ serves as the query, whereas the feature $\widehat{Z_l^{T\to Si}}$ serves as the key and value. We used MCA to inject temporal feature $\widehat{Z_l^{T\to Si}}$ into the spatial feature, then the result feature $\widetilde{Z_l^{Si\prime}}$ goes through the MSA and MLP to restructure the relationship between the joints, which can be formulated as Equation (6).

$$\begin{aligned}\widetilde{Z_l^{Si\prime}} &= \widehat{Z_{l-1}^{Si}} + MCA(LN(\widehat{Z_{l-1}^{Si}}), LN(\widehat{Z_{l-1}^{T\to Si}}), LN(\widehat{Z_{l-1}^{T\to Si}})),\\ \widetilde{Z_l^{Si\prime\prime}} &= \widetilde{Z_l^{Si\prime}} + MSA(LN(\widetilde{Z_l^{Si\prime}})),\\ \widehat{Z_l^{Si}} &= \widetilde{Z_l^{Si\prime\prime}} + MLP(LN(\widetilde{Z_l^{Si\prime\prime}})),\end{aligned} \tag{6}$$

where $l \in [1, \cdots, L_2]$ is the index of CDI layers. After these operations, we successfully injected the temporal features into the spatial domain and obtained the final spatial reconstruction features $\widehat{Z_{L_2}^{Si}}$ from the SR block following the L_2-layer CDI module.

3.3.2. Temporal Refinement

We propose that the TR utilizes the output of the SR to refine for temporal features that have weaker spatial structure relationships. TR consists of the FTU, MCA, and MLP (see Figure 4b). As in the previous operation in the SR, we need to convert the spatial features to the temporal domain. Each frame of the SR output $\widehat{Z_l^{Si}} \in \mathbb{R}^{J\times C}$ is reshaped into a vector $\overset{\frown}{Z_l^{Si}} \in \mathbb{R}^{1\times(J\cdot C)}$. Then, concatenating the N frames vectors $\overset{\frown}{Z_l^{S1}}, \overset{\frown}{Z_l^{S2}}, \cdots, \overset{\frown}{Z_l^{SN}}$ and changing the feature dimension form $\mathbb{R}^{N\times(J\cdot C)}$ to $\mathbb{R}^{N\times D}$ by a linear layer. Finally, the feature is normalized with the LayerNorm layer, denoted as $\widehat{Z_l^{S\to T}}$, and sent to MCA along with the temporal feature $\widehat{Z_l^T}$ to complement the weaker spatial structure relationship in the temporal domain. It differs from SR in that TR uses the temporal features as the query and the reconstructed spatial features as the key and value.

$$\begin{aligned}\widetilde{Z_l^{T\prime}} &= \widehat{Z_{l-1}^{T}} + MCA(LN(\widehat{Z_{l-1}^{T}}), LN(\widehat{Z_{l-1}^{S\to T}}), LN(\widehat{Z_{l-1}^{S\to T}})),\\ \widehat{Z_l^T} &= \widetilde{Z_l^{T\prime}} + MLP(LN(\widetilde{Z_l^{T\prime}})), l \in [1, \cdots, L_2].\end{aligned} \tag{7}$$

The output of the TR block in the last layer of the CDI module is $\widehat{Z_{L_2}^T}$.

3.4. Regression Head

In the regression head, our model learns two different linear regression functions to regress the 3D pose from the spatial and temporal domains, respectively. In the spatial domain, the spatial head is applied on $\widehat{Z_{L_2}^{Si}} \in \mathbb{R}^{J\times C}$ to map the dimension of each joint from C to 3 for generating the 3D pose $\widetilde{X^{S}}^{i} \in \mathbb{R}^{J\times 3}$ at frame i, where 3 denotes the joint coordinates in 3D space. The pose in the center frame is marked by $\widehat{X^S}$. For the temporal domain, the temporal head is applied on $\widehat{Z_{L_2}^T}$ to regress the 3D poses $\widetilde{X^T} \in \mathbb{R}^{N\times J\times 3}$. Then, the pose of the intermediate frame is selected and denoted as $\widehat{X^T} \in \mathbb{R}^{J\times 3}$. Ultimately, the final prediction of the model is the average of $\widehat{X^S}$ and $\widehat{X^T}$.

3.5. Loss Function

We employed the mean-squared error (MSE) as the loss function in our model, which is a common tool in 3D human pose estimation. The MSE optimizes the parameters of the model for better performance by minimizing the error between the predicted and ground truth positions.

Temporal loss:

$$L_T = \sum_{n=1}^{N} \sum_{j=1}^{J} \| Y_j^n - \widetilde{X^T}_j^n \|_2 \tag{8}$$

Spatial loss:

$$L_S = \sum_{n=1}^{N} \sum_{j=1}^{J} \| Y_j^n - \widetilde{X^S}_j^n \|_2 \tag{9}$$

where Y_j^n and $\widetilde{X}_j^n$ are the ground truth and estimated 3D positions of joint j at frame n, respectively.

The final loss of the whole model during the training stage:

$$L = \lambda_S L_S + \lambda_T L_T \tag{10}$$

where λ_S and λ_T are the weighting factors for spatial and temporal loss, respectively.

4. Experiments

The structure of this section is as follows. First, we give the dataset and evaluation metrics, followed by a description of our experimental setup and a comparison of STFormer with recent work. Finally, several ablation experiments are performed to demonstrate the superiority of our structure and the effectiveness of each module.

4.1. Dataset and Evaluation Metrics

Human3.6M [39] is a popular benchmark dataset for 3D human pose estimation. It contains 3.6 million human poses and corresponding images captured by four high-resolution 50 HZ cameras. A total of 15 actions were performed by 11 professional actors in indoor scenes, such as discussions, photo shoots, and smoking. As in previous work, our STFormer was selected for training on (S1, S5, S6, S7, S8) five objects and then tested on (S9, S11) two objects.

We evaluated the performance of STFormer on two commonly used evaluation metrics (MPJPE and P-MPJPE). The MPJPE is mean per joint position error, which calculates the Euclidean distance between the predicted joint position and the ground truth, also referred to as Protocol 1. The P-MPJPE calculates the MPJPE after aligning the prediction to the ground truth through rotation, translation, and scaling operations, referred to as Protocol 2.

4.2. Implementation Details

The proposed STFormer includes $L_1 = 2$ FE and $L_2 = 1$ CDI layers. We set the weighting factors λ_S and λ_T both equal to 0.5. We implemented STFormer using Python 3.7.12 under the Pytorch framework. A GeForce RTX 2080 Ti GPU was used to train and test the model. When training, we chose Amsgrad as the optimizer and terminated after 20 epochs. The initial learning rate and shrink factor were set to 0.001 and 0.95, respectively. For a fair comparison, we adopted the same augmentation strategy as in [8,14,15]. The 2D pose was obtained by a cascaded pyramid network (CPN) [40].

4.3. Results

4.3.1. Results on Human3.6M

A comparison of the results of our method with recent mainstream methods on the Human3.6M dataset is shown in Table 1. Table 1 lists the performance of STFormer with an input of nine frames under both metrics. Surprisingly, our method surpasses the previous state-of-the-art methods with respect to Protocol 1 (47.8 mm) and Protocol 2 (37.6 mm). In comparison to the Transformer-based network (PoseFormer), the STFormer improved by 2.1 mm on the MPJPE.

Table 1. Comparison of our method with previous methods on Human3.6M under Protocol 1 (top table) and Protocol 2 (bottom table) using the CPN-detected 2D pose as the input.

Protocol 1	**Dir**	**Disc**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Purch**	**Sit**	**SitD**	**Smoke**	**Wait**	**WalkD**	**Walk**	**WalkT**	**Avg**
Martinez et al. [29]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Fang et al. [41]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Hossain et al. [28]	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Lee et al. [42]	40.2	49.2	47.8	52.6	50.1	75.0	50.2	43.0	55.8	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Pavllo et al. [14]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.8
Cai et al. [15]	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
METRO [26]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	54.0
GraphSH [43]	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
PoseFormer [8]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49.9
MGCN [44]	45.4	49.2	45.7	49.4	50.4	58.2	47.9	46.0	57.5	63.0	49.7	46.6	52.2	38.9	40.8	49.4
STFormer (Ours)	43.6	47.3	45.3	46.5	48.3	56.5	44.7	43.3	57.3	66.9	48.4	44.7	49.9	35.6	38.6	47.8
Protocol 2	**Dir**	**Disc**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Purch**	**Sit**	**SitD**	**Smoke**	**Wait**	**WalkD**	**Walk**	**WalkT**	**Avg**
Martinez et al. [29]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	49.4	49.2	45.0	49.5	38.0	43.1	47.7
Fang et al. [41]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Hossain et al. [28]	35.7	39.3	44.6	43.0	47.2	54.0	38.3	37.5	51.6	61.3	46.5	41.4	47.3	34.2	39.4	44.1
Lee et al. [42]	34.9	35.2	43.2	42.6	46.2	55.0	37.6	38.8	50.9	67.3	48.9	35.2	50.7	31.0	34.6	43.4
Pavlakos et al. [25]	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Cai et al. [15]	35.7	37.8	36.9	40.7	39.6	45.2	37.4	34.5	46.9	50.1	40.5	36.1	41.0	29.6	33.2	39.0
Liu et al. [23]	35.9	40.0	38.0	41.5	42.5	51.4	37.8	36.0	48.6	56.6	41.8	38.3	42.7	31.7	36.2	41.2
STFormer (Ours)	33.5	36.9	36.2	37.6	37.7	43.3	34.0	32.9	46.4	52.9	39.1	34.1	39.4	27.5	31.7	37.6

4.3.2. Computation Complexity Analysis

Table 2 compares the computational complexity, MPJPE, and frames per second (FPS) with several previous methods on Human 3.6M. For the number of parameter analysis, our method was lower than the previous one with a significant increase in accuracy. Moreover, as can be seen in Table 3, the number of parameters of STFormer hardly increased with the number of receptive fields (f). Regarding the floating-point operations (FLOPs), the proposed method was not the best in comparison to the other methods. However, compared with the same Transformer-based approaches (PoseFormer [8] and CrossFormer [30]), our method was much lower in terms of FLOPs. Additionally, although the inference speed of the proposed model was lower than the other methods in Table 2, it still had an acceptable FPS for real-time inference [11].

Table 2. Comparison between the proposed method with a set of previous methods in terms of floating-point operations (FLOPs), number of the parameters, MPJPE, and frames per second (FPS). The experiments were conducted on Human3.6M under Protocol 1 with the CPN-detected 2D pose as the input.

Method	f	**Param (M)**	**FLOPs (M)**	**MPJPE**	**FPS**
Hossain et al. [28]	-	16.95	33.88	58.3	-
Pavllo et al. [14]	27	8.56	17	48.8	1492
PoseFormer [8]	9	9.58	150	49.9	320
CrossFormer [30]	9	9.93	163	48.5	284
STFormer (Ours)	9	7.21	137	47.8	155

Table 3. Results of ablation experiments with STFormer at different input frames.

Method	f	Param (M)	FLOPs (M)	MPJPE
STFormer	3	7.210262	45.9	50.6
STFormer	5	7.211286	76.6	49.5
STFormer	7	7.212310	107.2	48.6
STFormer	9	7.213334	137.8	47.8

4.4. Ablation Study

4.4.1. Impact of Receptive Fields

Low input video frames are very important for improving network efficiency and have higher real-world application value. To investigate the performance of STFormer under different low input frames, we conducted four comparative experiments, all using CPN as the extractor. Table 3 lists the outcomes of our approach with various input frames (f). From the results, it can be seen that the result was 50.8 with 3 input frames and the error rate was reduced by 2.8 when increasing from 3 to 9 frames. The table reveals that our technique can continue to deliver great performance at low input frames. In the following section of the ablation study, we uniformly took nine frames as input and evaluate under Protocol 1.

4.4.2. Structure Analysis

Our proposed STFormer is a two-branch structure with constant spatio-temporal communication. To illustrate the validity of our structure, we propose some structural variants and compared them with our approach. We can abstract our proposed structure as shown in Figure 5a. Under different information exchange mode, the CDI module can implement it in a summation way (as in Figure 5b). As seen in Table 4, the summation approach is inferior to our proposed method (difference of 0.9 mm). The explanation may be that our suggested strategy is able to utilize MCA adaptively across domains to obtain the required information. Figure 5b shows an undifferentiated summation process on the received information.

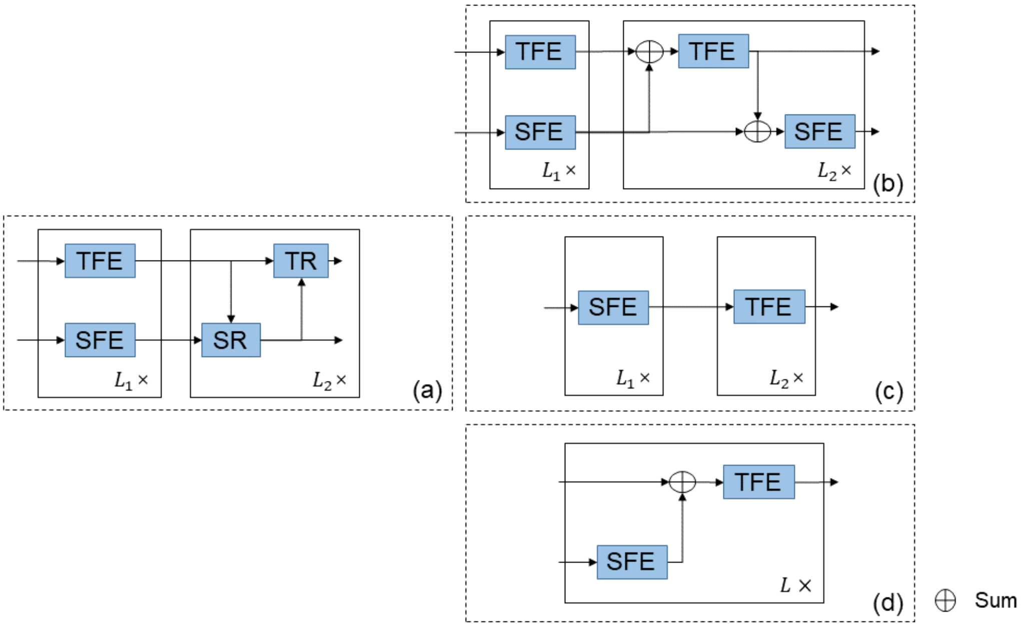

Figure 5. Structure analysis. (**a**) The abstracted model proposed in this paper. (**b**) A special case of replacing the MCA with a summation method in the CDI. (**c**) The structure of PoseFormer. (**d**) The structure of TNT. L, L_1, and L_2 are the number of layers of the corresponding module.

Furthermore, we explored the effect of temporal information, and we only transferred the spatial structure information to the temporal domain, as in Figure 5c,d. The structural design of PoseFormer [8] is used in Figure 5c, and the structural design of TNT [45] is used in Figure 5d. In the absence of injecting temporal features into the spatial domain, the results obtained in Table 4c,d are far below ours. This is a strong indication that temporal dependencies can help the model determine the state of motion in which the pose is in the current moment and, thus, adjust the structural relationships between joints, which is essential for accurate 3D human pose prediction. By comparing Table 4c,d, it is evident that frequent injection of unadjusted spatial features into the temporal domain is counterproductive and leads to performance degradation. This demonstrated that relying on Spatial Feature Extraction alone does not accurately model the spatial structure relationships of single-frame poses in the absence of temporal dependence information.

Table 4. Ablation study on various structures with the MPJPE (mm).

(a)	(b)	(c)	(d)
47.8	48.7	50.2	51.0

4.4.3. Impact of Each Component

In Table 5, we conducted experiments to confirm the contribution of each component. Initially, we exclusively used the features extracted by SFE or TFE to estimate the 3D pose of the intermediate frame (57.9 mm for SFE and 49.1 mm for TFE). The experimental results indicated that video can supply temporal dependency information to help the model obtain a more accurate 3D pose compared to a single frame. For the following comparison, we took the TFE results as our baseline. Then, we incorporated SFE and TFE with SR or TR. When combined with SR (FE-SR), the performance was significantly improved (by 0.9 mm), demonstrating the ability of the SR block to reconstruct spatial features. However, the performance did not improve when combined with TR (FE-TR) only, suggesting that unadjusted spatial features do not effectively complement the weaker spatial information in the temporal domain. These ablation experiments showed that the SR block is indeed capable of adjusting the spatial structure between joints using temporal information and further refining the temporal domain features, which are of high value for 3D human pose estimation. Finally, we combined all the proposed modules and obtained a result of 47.8, which reduced the error by 1.3 compared to the baseline.

Table 5. Validation of the effectiveness of the different components.

Method	SFE	TFE	SR	TR	MPJPE
SFE	✓	×	×	×	57.9
TFE	×	✓	×	×	49.1
FE-SR	✓	✓	✓	×	48.2
FE-TR	✓	✓	×	✓	49.1
STFormer	✓	✓	✓	✓	47.8

4.4.4. Parameter Setting Analysis

Table 6 illustrates how varying hyper-parameter settings affect the performance of our approach. The network has four main hyper-parameters: the depth of FE (L_1), the depth of CDI (L_2), the dimension of the spatial domain (d_S), and the dimension of the temporal domain (d_T). To analyze the impact and selection of each configuration, we assigned a different value to one of the hyper-parameters while keeping the others the same. It is shown that expanding the embedding dimension from 16 to 32 in the spatial domain or from 256 to 512 in the temporal domain can boost the performance, but no further improvement was achieved when the number of dimensions continued to increase. In addition, we observed that the best results were obtained when $L_1 = 2$ and $L_2 = 1$, while continuing to stack more layers brought no performance improvement. As the final network parameters, we selected the combination $L_1 = 2$, $L_2 = 1$, $d_S = 32$, and $d_T = 512$ based on this table.

Table 6. Ablation experiments for different parameter settings of our model.

L_1	L_2	d_S	d_T	**MPJPE**
2	1	16	512	49.7
2	1	32	512	47.8
2	1	48	512	48.8
2	1	32	256	48.9
2	1	32	512	47.8
2	1	32	768	48.1
2	1	32	512	47.8
3	1	32	512	48.9
2	2	32	512	48.7
1	2	32	512	49.0

4.5. Qualitative Results

The visualization results of the STFormer and baseline method are shown in Figure 6 with an input of nine frames. We used the green arrows to point out the locations of the prediction errors of the baseline method. As can be seen from Figure 6, the method proposed in this paper clearly outperformed the baseline method. In addition, we tested on some challenging natural scenes, and the results are shown in Figure 7. These images from natural scenes rarely appear in the Human3.6M dataset. For these challenging scenes, our method was still able to predict the 3D pose accurately.

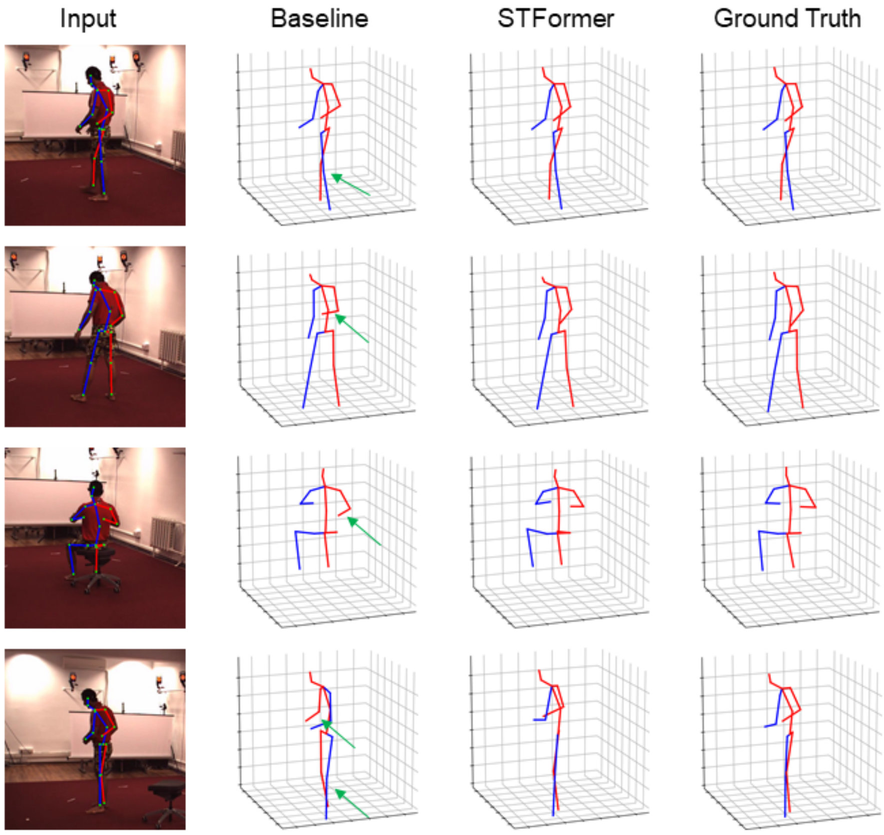

Figure 6. Qualitative comparison of the proposed method (STFormer) with the baseline method on the Human3.6M dataset. Incorrect estimates are highlighted with green arrows.

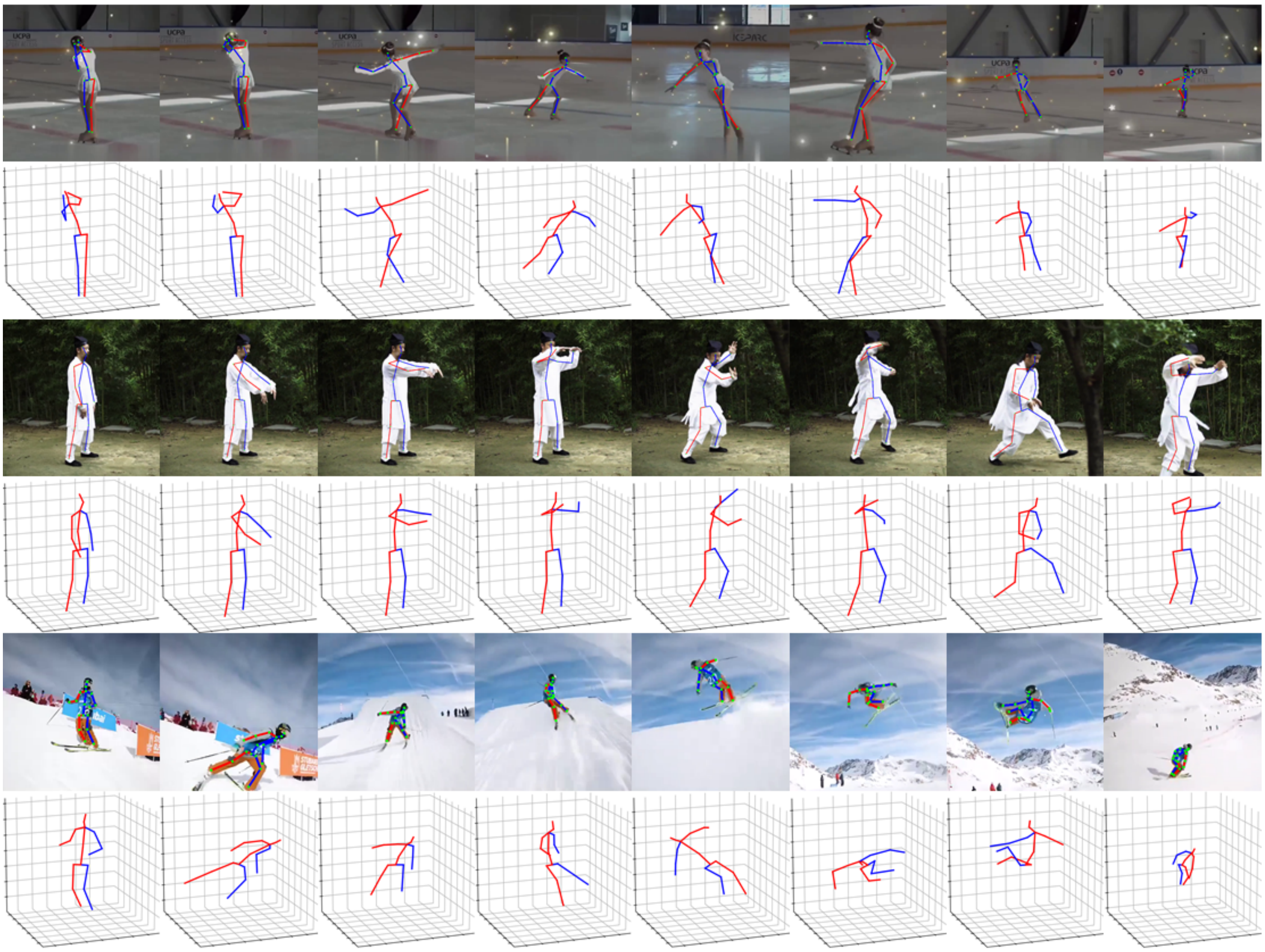

Figure 7. Visualization results of the proposed STFormer for reconstructing 3D human poses on challenging natural scene images.

5. Conclusions

In this work, we proposed STFormer, a two-stage network for estimating 3D human poses from monocular video based on Transformer. Unlike previous sequential modeling approaches that first extract spatial structural features and then learn temporal consistency, STFormer not only extracts spatio-temporal features, but also considers the effect between information from different domains (temporal and spatial domains). The model consists of two stages: Feature Extraction (FE) and Cross-Domain Interaction (CDI). In FE (Stage 1), we proposed two similar blocks, SFE and TFE, for extracting the features from the spatial and temporal domains, respectively. However, there was no interaction between these two features. To enhance the representation of the current domain by utilizing features from other domains, we proposed SR and TR in the CDI module (Stage 2). The SR block injects temporal information into the spatial domain to adjust the spatial structure relationships, and then, the output of SR is fed to TR to compensate for the weaker spatial structure relationships of the temporal features. Detailed experiments demonstrated that the proposed STFormer has a basic benefit over purely spatial or temporal Transformer and achieved better performance than mainstream methods on benchmark datasets.

Author Contributions: Methodology, H.W.; software, H.W.; validation, H.W., B.S. and Q.S.; formal analysis, H.W.; investigation, Q.S., H.W. and B.S.; writing—original draft preparation, H.W.; writing—review and editing, H.W., Q.S. and B.S.; supervision, Q.S.; funding acquisition, Q.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by The Natural Science Foundation of Hebei Province (F2019201451).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Not applicable

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, M.; Liu, H.; Chen, C. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognit.* **2017**, *68*, 346–362. [CrossRef]
2. Liu, M.; Yuan, J. Recognizing human actions as the evolution of pose estimation maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1159–1168.
3. Wang, P.; Li, W.; Gao, Z.; Tang, C.; Ogunbona, P.O. Depth pooling based large-scale 3-d action recognition with convolutional neural networks. *IEEE Trans. Multimed.* **2018**, *20*, 1051–1061. [CrossRef]
4. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
5. Errity, A. Human–computer interaction. In *An Introduction to Cyberpsychology*; Routledge: Milton Park, UK, 2016; pp. 263–278.
6. Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.P.; Xu, W.; Casas, D.; Theobalt, C. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph.* **2017**, *36*, 1–14. [CrossRef]
7. Zheng, J.; Shi, X.; Gorban, A.; Mao, J.; Song, Y.; Qi, C.R.; Liu, T.; Chari, V.; Cornman, A.; Zhou, Y.; et al. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4478–4487.
8. Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; Ding, Z. 3d human pose estimation with spatial and temporal transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11656–11665.
9. Wang, J.; Yan, S.; Xiong, Y.; Lin, D. Motion guided 3d pose estimation from videos. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 764–780.
10. Chen, T.; Fang, C.; Shen, X.; Zhu, Y.; Chen, Z.; Luo, J. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 198–209. [CrossRef]
11. Li, W.; Liu, H.; Ding, R.; Liu, M.; Wang, P.; Yang, W. Exploiting temporal contexts with strided Transformer for 3d human pose estimation. *IEEE Trans. Multimed.* **2022**, *25*, 1282–1293. [CrossRef]
12. Iskakov, K.; Burkov, E.; Lempitsky, V.; Malkov, Y. Learnable triangulation of human pose. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7718–7727.
13. Qiu, H.; Wang, C.; Wang, J.; Wang, N.; Zeng, W. Cross view fusion for 3d human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4342–4351.
14. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7753–7762.
15. Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.J.; Yuan, J.; Thalmann, N.M. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2272–2281.
16. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual attention network. *arXiv* **2022**, arXiv:2202.09741.
17. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-aware trident networks for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6054–6063.
18. Kumar, D.; Kukreja, V. Early Recognition of Wheat Powdery Mildew Disease Based on Mask RCNN. In Proceedings of the 2022 International Conference on Data Analytics for Business and Industry (ICDABI), online, 25–26 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 542–546.
19. Kumar, D.; Kukreja, V. MRISVM: A Object Detection and Feature Vector Machine Based Network for Brown Mite Variation in Wheat Plant. In Proceedings of the 2022 International Conference on Data Analytics for Business and Industry (ICDABI), Online, 25–26 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 707–711.
20. Kumar, D.; Kukreja, V. Application of PSPNET and Fuzzy Logic for Wheat Leaf Rust Disease and its Severity. In Proceedings of the 2022 International Conference on Data Analytics for Business and Industry (ICDABI), online, 25–26 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 547–551.

21. Kumar, D.; Kukreja, V. A Symbiosis with Panicle-SEG Based CNN for Count the Number of Wheat Ears. In Proceedings of the 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 13–14 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–4.
22. Zeng, A.; Sun, X.; Yang, L.; Zhao, N.; Liu, M.; Xu, Q. Learning skeletal graph neural networks for hard 3d pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11436–11445.
23. Liu, K.; Ding, R.; Zou, Z.; Wang, L.; Tang, W. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part X 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 318–334.
24. Moon, G.; Lee, K.M. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part VII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 752–768.
25. Pavlakos, G.; Zhou, X.; Daniilidis, K. Ordinal depth supervision for 3d human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7307–7316.
26. Lin, K.; Wang, L.; Liu, Z. End-to-end human pose and mesh reconstruction with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1954–1963.
27. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-fine volumetric prediction for single-image 3D human pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7025–7034.
28. Hossain, M.R.I.; Little, J.J. Exploiting temporal information for 3d human pose estimation. In Proceedings of the European conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 68–84.
29. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A simple yet effective baseline for 3d human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2640–2649.
30. Hassanin, M.; Khamiss, A.; Bennamoun, M.; Boussaid, F.; Radwan, I. CrossFormer: Cross spatio-temporal Transformer for 3d human pose estimation. *arXiv* **2022**, arXiv:2203.13387.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
32. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based object re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15013–15022.
33. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision Transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
34. Yang, G.; Tang, H.; Ding, M.; Sebe, N.; Ricci, E. Transformer-based attention networks for continuous pixel-wise prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 16269–16279.
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
36. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part I 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
37. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.T.; Zhou, E. Tokenpose: Learning keypoint tokens for human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11313–11322.
38. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
39. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]
40. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
41. Fang, H.S.; Xu, Y.; Wang, W.; Liu, X.; Zhu, S.C. Learning pose grammar to encode human body configuration for 3d pose estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
42. Lee, K.; Lee, I.; Lee, S. Propagating lstm: 3d pose estimation based on joint interdependency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 119–135.
43. Xu, T.; Takano, W. Graph stacked hourglass networks for 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16105–16114.

44. Zou, Z.; Tang, W. Modulated graph convolutional network for 3D human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11477–11487.
45. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in Transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.